THE IDENTIFIABILITY OF TREE TOPOLOGY FOR PHYLOGENETIC MODELS, INCLUDING COVARION AND MIXTURE MODELS

ELIZABETH S. ALLMAN AND JOHN A. RHODES

Abstract. For a model of molecular evolution to be useful for phylogenetic inference, the topology of evolutionary trees must be identifiable. That is, from a joint distribution the model predicts, it must be possible to recover the tree parameter.

We establish tree identifiability for a number of phylogenetic models, including a covarion model and a variety of mixture models with a limited number of classes.

The proof is based on the introduction of a more general model, allowing more states at internal nodes of the tree than at leaves, and the study of the algebraic variety formed by the joint distributions to which it gives rise. Tree identifiability is first established for this general model through the use of certain phylogenetic invariants.

1. Introduction

In phylogenetics, probabilistic models of the evolution of biological sequences (DNA or proteins, for example) are used to infer evolutionary history.

The parameters of such a model typically include such things as the topology of the rooted tree depicting the temporal ordering of speciation events, the elapsed time between these events, and the rates at which different types of substitutions ($A \to C$, $A \to G$, etc.) occur between events. While any of these parameters might be of interest in a particular study, the tree topology is often the one of greatest interest (and one on which the very definition of the others depends).

A basic question concerning any statistical model then is whether it is *identifiable*: Given a distribution of observations that the model predicts, is it theoretically possible to recover the parameters of the model? Understanding what parameters are identifiable for a model is crucial to understanding what we may reasonably hope to infer from data.

Date: November 7, 2005.

In particular, identifiability of the tree topology is essential for any model that is to be used in inferring evolutionary histories from data. If a tree is not uniquely determined by an expected joint distribution, then one has no hope of using the model to infer trees well from data. Indeed, proofs of the statistical consistency of an inference method such as maximum likelihood begin by establishing identifiability of parameters.

Identifiability of tree topologies has been investigated for a number of models. See, for example, [CH91, SSH94, Cha96, WS97, Baa98, SHP98, Rog01] for both positive and negative results. However, much remains to be done. As pointed out in [BGP05], for mixture models that allow several classes of sites in sequences to evolve at different rates, "[n]othing has been proved in the general context yet." In fact, most proofs of tree identifiability for various models have been based on notions of phylogenetic distances and the four-point condition of Buneman [Bun71], and for general mixture models no such distance is known. (See also [EW04] for related work illustrating non-identifiability of edge lengths for mixture models.)

In this paper we establish a general identifiability result applicable to many mixture models, albeit with a limited number of classes. As further motivation for our work, however, we choose to highlight the *covarion model*, for which the question of identifiability of trees has also been open.

A covarion model of character evolution describes characters with states that are only partially observable. Such a model can be viewed as a type of hidden Markov model on a tree, where not only are character states at all internal nodes of the tree hidden, as in simpler phylogenetic models, but even at the leaves full information is not available.

For instance, in the covarion model of Tuffley and Steel [TS98] for the evolution of DNA sequences, each site in the sequences is a character. As evolution proceeds over a tree, this character is in one of the eight states A_{on}, A_{off}, G_{on}, G_{off}, C_{on}, C_{off}, T_{on}, T_{off}, where the subscript *on* denotes the site is currently free to undergo base substitution, and *off* denotes that it is currently invariable. The off-states indicate functional or other biological constraints temporarily preventing substitutions. Importantly, as evolution proceeds over the tree, sites may pass from on-states to off-states and *vice versa*. However, when we observe sequences from currently extant taxa, we can only observe A, C, G, or T; we obtain no information as to whether a site is currently on or off.

That constraints to nucleotide substitution might change over a tree is a biologically plausible hypothesis that makes covarion models attractive. A covarion model might be viewed as a type of rate variation model, as several characters described by the same model may be in on- and off-states for different durations, and thus undergo different amounts of substitution. However, the 'switching' between on- and off-states, allowing a character to behave differently in one part of a tree from another, is a crucial distinction from standard approaches to rate variation. Of course more elaborate covarion models, with more than the two hidden rate-classes of the example above, are easily devised. For instance there might be off-, slow-, and fast-states, with the opportunity for a character to pass in and out of each as evolution proceeds over the tree.

Though originally proposed by Fitch and Markowitz [FM70], it was not until the work of Tuffley and Steel [TS98] that a covarion model was mathematically formalized and the first steps were taken in its theoretical analysis. More recently, Galtier [Gal01] implemented a maximum likelihood inference package using a covarion model, and reported improved fits to data over standard rate-variation models. See also [PMCH01] for a more thorough overview and arguments in support of the use of such models.

Although the covarion model is appealing for biological reasons, it is less well understood theoretically. For instance, many basic questions of identifiability of model parameters have been open. Indeed, much of [TS98] is focused on showing that in some circumstances a covarion model is distinguishable from a rate-variation model, and that some features of the tree topology are identifiable provided one has prior knowledge of some clades.

Motivated by the covarion model, in this paper we first prove a result on identifiability of tree topologies for a more general phylogenetic model. The model is introduced in Section 2, and the result proved in Section 4. We also show how the result specializes to establish identifiability of the tree topology for more specific models of greater direct interest for applications, including the covarion model and certain mixture models. In Sections 3 we describe some of these models, and in Section 5 we apply the general result to deduce tree identifiability under certain assumptions.

Actually, our results require some mild restrictions on model parameters — it is better to say that tree topology is identifiable for *generic* parameters. Informally this means if parameters are chosen "at random," then the topology can be identified. More precisely, our use of

the word "generic" is as in algebraic geometry: We say a property holds for generic parameters if it holds for all parameters off of a proper subvariety of the parameter space. By "subvariety" we might mean either an algebraic subvariety, defined by the vanishing of a set of multivariable polynomials, or, more generally, an analytic subvariety, defined by the vanishing of analytic functions. Since in either circumstance a proper subvariety is a closed set of lower dimension than the ambient space, generic parameters form an open, dense subset of the parameter space.

As the last paragraph hints, our approach throughout is algebraic, and provides a good illustration of the value of an algebraic viewpoint for statistical models. Within phylogenetics, this approach began with the introduction of the idea of *phylogenetic invariants* by Cavender and Felsenstein [CF87] and Lake [Lak87]. Notable contributions for group-based models appeared in [ES93] and [SSE93], building on an idea first introduced in [Hen89]. It has been pursued in a number of recent works focused on phylogenetics, such as [CHHP00, AR03, CKS03, AR04, ERSS04, SS05, AR05c, AR05b, CGS05, CS05, Eri05], and more broadly for biological application in the recent volume [PS05].

Though our emphasis here is on theory, practical methods of identifying tree topologies from data are also needed. For instance, the notions of phylogenetic distance that play a key role in theoretical identification of tree topologies for simpler models also provide useful tools for tree inference. Whether one wishes to base inference on a distance-based method, or merely view such methods as fast heuristic means of finding good candidate trees to begin a more elaborate search of tree space, the value of distances is clear. For models where no distance formulas are known, the explicit polynomials our results yield, whose vanishing on a joint distribution identifies the tree topology, might play a similar role. It will be interesting to see if these polynomials might be exploited for practical inference, either heuristically or on a more solid statistical basis.

Finally, we thank Cecile Ané for first suggesting to us that the covarion model might be tractably studied by our methods.

2. The (λ, κ)-state general Markov model

In this section we introduce a phylogenetic model which allows more states at internal nodes of the tree than at leaves. Though motivated by the covarion model of [TS98], our model is much more general. We emphasize that we introduce this model not because we feel it precisely

captures any biological phenomena, but rather because its generality encompasses a variety of models of more direct biological interest. It will allow us to make the key algebraic ideas in our subsequent arguments on identifiability clear, and obtain results which can then be applied to more specialized models.

Throughout, suppose T is a trivalent (i.e., binary) tree. Choosing some *internal* vertex r as a root, denote the resulting rooted tree by T^r. Corresponding to each leaf of the tree we have an observed random variable with state space $[\kappa] = \{1, 2, \ldots, \kappa\}$, while for each internal vertex we have a hidden (unobserved) variable with state space $[\lambda]$. The states of observed variables might represent the bases at a site in DNA sequences ($\kappa = 4$) from extant taxa, while states of hidden variables might represent ancestral bases together with additional features, such as how rapidly a site currently undergoes mutation, or even whether it is currently invariable. With this interpretation in mind, we will always assume $\lambda \geq \kappa$.

A λ-element row vector $\boldsymbol{\pi}_r$ describes the probability distribution of the states for the root variable. For each internal edge e of the tree, with e directed away from the root, a $\lambda \times \lambda$ Markov matrix M_e describes transition probabilities. For each pendant edge e, a $\lambda \times \kappa$ Markov matrix M_e describes transition probabilities. Thus $M_e(i, j)$ is the conditional probability of state j at the end of e given state i at its start. Stochastic assumptions ensure that all entries are non-negative, the entries of $\boldsymbol{\pi}_r$ sum to 1, and the entries in any row of any of the M_e sum to 1. With no further restrictions imposed on either the root distribution or the Markov matrices, we call this the (λ, κ)-*state general Markov model* on the rooted tree T^r.

In the case $\lambda = \kappa$, this model is the usual general Markov model with κ states. We are therefore particularly interested in cases where $\lambda > \kappa$. For instance, a generalization of the on-off covarion model described in the introduction has $\lambda = 2\kappa$.

For a fixed n-leaf rooted tree T^r, we may make some choice of entries in $\boldsymbol{\pi}_r$ and each row of the M_e to view as independent variables, using the condition that rows sum to 1 to determine the remaining entry. Since T^r has n pendant edges and $n - 3$ internal edges, the stochastic parameter space S for this model can thus be identified with a subset of $[0, 1]^L$ where $M = (\lambda - 1) + n\lambda(\kappa - 1) + (n - 3)\lambda(\lambda - 1)$.

The probabilities of observing each of the κ^n possible patterns (i.e., assignments of states) of leaf variables can then be given as polynomial expressions in the parameters. That is, there is a polynomial map, the

parameterization map,

$$\phi_{T^r} : S \to [0,1]^{\kappa^n},$$

which gives the joint distribution of observed states at the leaves of T^r as a function of the parameters. We extend this to a polynomial map $\mathbb{C}^L \to \mathbb{C}^{\kappa^n}$ which we also denote by ϕ_{T^r}, and refer to $\mathbb{C}^L$ as the *complex parameter space.*

The *phylogenetic variety* for the (λ, κ)-state model on T^r is the algebraic variety defined as

$$V_{T^r,\lambda,\kappa} = \overline{\phi_{T^r}(\mathbb{C}^L)},$$

where the bar denotes the (Zariski and standard) topological closure in $\mathbb{C}^{\kappa^n}$.

Lemma 1. *Let T be an n-leaf trivalent tree, and r_1, r_2 any two internal nodes. Then $V_{T^{r_1},\lambda,\kappa} = V_{T^{r_2},\lambda,\kappa}$, so this variety may be denoted by $V_{T,\lambda,\kappa}$.*

Proof. This is proved similarly to the corresponding result for the general Markov model, as in [SSH94] or[AR03]. □

3. Algebraic and analytic submodels of the (λ, κ)-state general Markov model

To further motivate our introduction of the (λ, κ)-state general Markov model, we note that many models of molecular evolution can be viewed as submodels of it, in that they simply place more restrictive assumptions on the allowable parameter values. In this section we first indicate some of these submodels of interest.

We also introduce the idea of an *analytic (λ, κ)-state model*, which will allow us not only to deal with parameterization maps in which joint distributions are expressed by polynomial formulas in the parameters, but a wider class that encompasses the 'rate matrix' models that are so commonly used in applications.

Some specific examples of models that can be viewed as submodels of the general (λ, κ)-state model include the following:

(1) GM: As already stated, the κ-state general Markov model results from $\lambda = \kappa$.
(2) GM+I: This model allows two classes of sites in sequences; one class mutates according to the general Markov model, and another is held invariable. A parameter f denotes the proportion of sites in the first class, with $1 - f$ in the second. If the root distribution vectors for the two classes are $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, let $\boldsymbol{\pi}_r = (f\boldsymbol{\pi}_1, (1-f)\boldsymbol{\pi}_2)$. For an internal edge e, if N_e is the

$\kappa \times \kappa$ matrix describing transition probabilities for the first class along that edge, let

$$M_e = \begin{pmatrix} N_e & 0 \\ 0 & I \end{pmatrix},$$

a $2\kappa \times 2\kappa$ matrix. For pendent edges e, let

$$M_e = \begin{pmatrix} N_e \\ I \end{pmatrix},$$

a $2\kappa \times \kappa$ matrix. Thus the model results from simply restricting parameters in the general $(2\kappa, \kappa)$-state model so that all Markov matrices have a particular form.

The restricted parameter space can be identified with $\mathbb{C}^M$, where $M = (2\kappa - 1) + (2n-3)\kappa(\kappa - 1)$. Note the parameterization map giving joint distributions as a function of parameters for this model is still a polynomial one, given by restriction of the map for the general $(2\kappa, \kappa)$-state model to a smaller domain.

(3) GM+GM+$\cdots$ +GM: We consider m classes of sites, each evolving independently according to a different GM model. To view this as a submodel of our more general model, for each internal edge of the tree we create an $m\kappa \times m\kappa$ block-diagonal Markov matrix, with each of the m $\kappa \times \kappa$ blocks giving transition probabilities for a particular class. On pendant edges we 'stack' the blocks, giving $m\kappa \times \kappa$ matrices. The root distribution is similarly obtained by concatenating the root distributions for each class, weighted by additional parameters describing the relative frequencies of the classes. Thus we are dealing with a restriction of the $(m\kappa, \kappa)$-state model, with parameter space identified with $\mathbb{C}^M$ where $M = (m\kappa - 1) + (2n-3)m\kappa(\kappa - 1)$. Again, the parameterization map for this model is polynomial.

(4) Other algebraic models: In the previous examples, we can replace an occurrence of GM by a submodel, such as the Jukes-Cantor, Kimura 2-parameter, Kimura 3-parameter, or Strand Symmetric, defined by further restriction of parameters. Allowing arbitrary matrices of these types on each edge (so that we do *not* assume a common rate matrix), we again have a polynomial parameterization map with domain $\mathbb{C}^M$ for some M.

The previous examples all lie fully within an algebraic framework, but in fact many of the models used for inference in current applications are not of this sort. *Rate matrix models* assume more commonality to the substitution process on the various edges of the tree. Typically, one fixes a rate matrix Q with non-negative off-diagonal entries and

rows summing to 0. Then each edge of the tree is assigned a scalar edge length t_e, and the Markov matrix $M_e = \exp(Qt_e)$ gives transition probabilities for that edge.

(5) GTR: A submodel of the general (κ, κ)-state model, the general time-reversible model assumes a root distribution $\boldsymbol{\pi}_r$ and rate matrix Q such that $\boldsymbol{\pi}_r Q = \mathbf{0}$ and $\mathrm{diag}(\boldsymbol{\pi}_r)Q$ is symmetric. Pairs $\boldsymbol{\pi}_r, Q$ with these properties can be parameterized by $(\kappa - 1) + (\kappa)(\kappa-1)/2$ scalars. Since we may normalize so one edge length is 1, the parameter space for the full model is of dimension $M = (\kappa - 1) + \kappa(\kappa - 1)/2 + (2n - 4)$. However, the parameterization map giving joint distributions is not polynomial, as it involves a composition of matrix exponentials with the general Markov parameterization. Nonetheless, it is an analytic map.

(6) GTR+rate-classes: Let $\boldsymbol{\pi}_r, Q$ be as in the GTR model. Assuming m different classes of sites, we assign each a relative frequency f_i and a scalar rate parameter λ_i, with $\lambda_1 = 1$. Then the ith class undergoes substitutions on an edge e according to $N_e = \exp(Q\lambda_i t_e)$. For internal edges of the tree we embed the N_e as blocks in a larger block-diagonal matrix M_e, while for pendant edges we stack them, obtaining an expression of this model as a submodel of the general $(m\kappa, \kappa)$-state model. The parameter space is of dimension $M = (\kappa - 1) + \kappa(\kappa - 1)/2 + 2(m - 1) + (2n - 4)$. Again we have an analytic, but not polynomial, parameterization map.

Note that our formulation requires a finite number of rate classes. While current literature often refers to a continuous distribution of rates (usually with a Γ distribution), in practice inference is always done with a discretization of the distribution, producing finitely many rate classes as here.

(7) GTR+I+rate-classes: This model is simply the last, with the further assumption $\lambda_2 = 0$.

(8) Covarion: As formulated in [TS98], the Tuffley-Steel covarion model hypothesizes a common $2\kappa \times 2\kappa$ rate matrix Q of a particular form. Let $\boldsymbol{\pi}, R$ be a root distribution and rate matrix for the κ-state GTR model. Let $s_1, s_2 > 0$, and set $\sigma_1 = \frac{s_2}{s_1+s_2}$, $\sigma_2 = \frac{s_1}{s_1+s_2}$. Then

$$Q = \begin{pmatrix} R - s_1 I & s_1 I \\ s_2 I & -s_2 I \end{pmatrix}$$

is the rate matrix for a 2κ state time-reversible process, stationary on the root distribution vector $\boldsymbol{\pi}_r = (\sigma_1\boldsymbol{\pi}, \sigma_2\boldsymbol{\pi})$. In the

language of the introduction, the first κ states represent bases that are on, and the last κ ones that are off.

The covarion model then associates to each internal edge e of the tree the Markov matrix $M_e = \exp(Qt_e)$, and to a pendant edge e the matrix $M_e = \exp(Qt_e)(I_{\kappa\times\kappa}\ I_{\kappa\times\kappa})^T$. This model is therefore a submodel of the $(2\kappa, \kappa)$-state model.

Since the parameters for the covarion model can be viewed as $(\boldsymbol{\pi}, R, s_1, s_2, \{t_e\})$, the parameter space is of dimension $M = (\kappa-1)+\kappa(\kappa-1)/2+2+(2n-4)$. As with all rate matrix models, the parameterization map is analytic, though not polynomial.

(9) The model referred to as the SSRV in [Gal01] generalizes the covarion model to m rate classes, sharing the same rate matrix R, with switching allowed between the classes. It can similarly be seen to be a submodel of the $(m\kappa, \kappa)$-state model, with an analytic parameterization map.

Of course many more variations are possible. The reader familiar with other basic models will have no trouble modifying the examples above, adding rate classes if desired, or even mixing several different models as separate classes.

To formalize a notion of submodel of the (λ, κ)-state model that encompasses the above and other examples, we introduce some new terminology.

By a *submodel* of the (λ, κ)-state general Markov model on a tree T^r we mean a restriction of parameters to a subset of the full parameter space $\mathbb{C}^L$. Suppose the set of (λ, κ)-state general Markov model parameters $s \in \mathbb{C}^L$ allowed in the submodel is $\psi(U)$, the image under some analytic map $\psi : U \to \mathbb{C}^L$ of an open set $U \subseteq \mathbb{R}^M$. Then we say the submodel is an *analytic* (λ, κ)*-state model* with *parameter space* U and *Markov map* ψ. The *parametrization map* for the analytic model is then $\phi_{T^r} \circ \psi$, where ϕ_{T^r} is the parameterization map for the general (λ, κ)-state model:

$$U \xrightarrow{\psi} \mathbb{C}^L \xrightarrow{\phi_T^r} \mathbb{C}^{\kappa^n}.$$

Thus, for instance, the covarion model is a analytic $(2\kappa, \kappa)$-state model, as is the GM+I model. Note that algebraic submodels, with polynomial Markov maps, are included among the analytic ones. Analyticity of the Markov map will be important for our arguments in Section 5.

While all of the enumerated models above are analytic (λ, κ)-state models, they in fact have additional structure in common. First note that in each $\lambda = m\kappa$ for some m. Moreover, the set of states $[\lambda]$ at each

internal node is naturally identified with the set $[\kappa] \times [m]$. Under this identification, if we refer to the observable states in $[\kappa]$ as 'bases', then a state (i, j) represents 'base i' and 'class j'. Here 'class' might refer to 'rate class', or some other characteristic, such as the on/off feature in the covarion model. Finally, in all of these models the Markov matrices on pendant edges of the tree have a form

$$\widetilde{M} = M(I\ I\ \ldots\ I)^T,$$

where M is an $m\kappa \times m\kappa$ Markov matrix of the sort allowed on internal edges. Essentially this means the model hides all class information, so only bases are observable at leaves. We refer to such an analytic $(m\kappa, \kappa)$-state model as an *analytic κ-base, m-class model*.

4. Identifiability of the tree topology for the general (λ, κ)-state model

Returning to consideration of the (λ, κ)-state general Markov model, in this section we establish our main technical result on generic identifiability of tree topologies.

We first consider identifiability of the tree topology from a joint distribution of states at the leaves in the case of a 4-leaf tree. Let the three possible trivalent trees with leaves a, b, c, d be denoted by

$$T_1 = T_{ab|cd},\ T_2 = T_{ac|bd},\ T_3 = T_{ad|bc},$$

where the subscript $uv|wx$ denotes leaves u, v are adjacent to a common internal node, as are leaves w, x.

Focusing on T_1, denote the internal nodes by r, f, so that the root r is adjacent to the leaves a and b. For $s \in \mathbb{C}^L$, the complex parameter space, let the corresponding vector and matrix parameters (with rows summing to 1) be $\boldsymbol{\pi}_r \in \mathbb{C}^\lambda$, $M_{rf} \in M_{\lambda\times\lambda}(\mathbb{C})$, $M_{ra}, M_{rb}, M_{fc}, M_{fd} \in M_{\lambda\times\kappa}(\mathbb{C})$.

Then $P = \phi_{T_1}(s)$ can be expressed as a $\kappa \times \kappa \times \kappa \times \kappa$ tensor whose entries $P(i, j, k, l)$ give the 'probability' of observing states i, j, k, l at leaves a, b, c, d, respectively, and are given by the following formula:

Let $A = \mathrm{diag}(\boldsymbol{\pi}_r)M_{rf}$, a $\lambda \times \lambda$ matrix. Then define a $\lambda \times \lambda \times \lambda \times \lambda$ tensor Q by

$$Q(i, j, k, l) = \begin{cases} A(i, k) & \text{if } i = j \text{ and } k = l, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, let each of the $M_{ra}, M_{rb}, M_{fc}, M_{fd}$ act in the consecutive indices of Q to yield P, i.e.,

$$P(i,j,k,l) = \sum_{i',j',k',l'=1}^{\lambda} Q(i',j',k',l')M_{ra}(i',i)M_{rb}(j',j)M_{fc}(k',k)M_{fd}(l',l). \tag{1}$$

To motivate Equation (1), and our subsequent arguments, one should think of the matrix A as expressing the joint distribution of states at the vertices r and f. The tensor Q then represents the joint distribution for a (λ,λ)-state model on a quartet tree where no state changes occur along the pendant edges (i.e., the Markov matrices on these edges are I), as illustrated at the left in Figure 1. The model producing P 'extends' these terminal edges, placing the $\lambda \times \kappa$ Markov matrices on the extensions, as shown on the right.

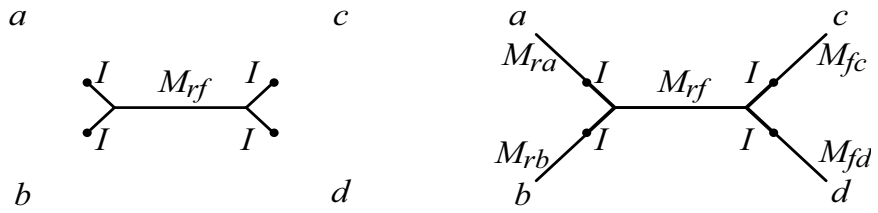

FIGURE 1. Models underlying the tensors Q and P.

Importantly, Equation (1) can be re-expressed several ways. The first of these naturally reflects the topology of T_1. Let $\mathrm{Flat}_{ab|cd}(Q)$ be the $\lambda^2 \times \lambda^2$ matrix with entries

$$\mathrm{Flat}_{ab|cd}(Q)((i,j),(k,l)) = Q(i,j,k,l),$$

where each index runs through $[\lambda]^2$. Similarly let $\mathrm{Flat}_{ab|cd}(P)$ be the $\kappa^2 \times \kappa^2$ matrix, with indices in $[\kappa]^2$, given by

$$\mathrm{Flat}_{ab|cd}(P)((i,j),(k,l)) = P(i,j,k,l).$$

Let $N_{ab} = M_{ra} \otimes M_{rb}$ and $N_{cd} = M_{fc} \otimes M_{fd}$ where '$\otimes$' denotes the Kronecker product of matrices. Thus N_{ab} and N_{cd} are $\lambda^2 \times \kappa^2$ matrices where, for instance $N_{ab}((i,j),(k,l)) = M_{ra}(i,k)M_{rb}(j,l)$. Then we have

$$\mathrm{Flat}_{ab|cd}(P) = N_{ab}^T \,\mathrm{Flat}_{ab|cd}(Q) N_{cd}. \tag{2}$$

Alternatively, we have other expressions involving flattenings which are less natural with respect to the topology of T_1. Let $\mathrm{Flat}_{ac|bd}(Q)$ be

the $\lambda^2 \times \lambda^2$ matrix with entries

$$\mathrm{Flat}_{ac|bd}(Q)((i,j),(k,l)) = Q(i,k,j,l).$$

Similarly let $\mathrm{Flat}_{ac|bd}(P)$ be the $\kappa^2 \times \kappa^2$ matrix with entries

$$\mathrm{Flat}_{ac|bd}(P)((i,j),(k,l)) = P(i,k,j,l).$$

Let $N_{ac} = M_{ra} \otimes M_{fc}$ and $N_{bd} = M_{rb} \otimes M_{fd}$. Then we have

$$\mathrm{Flat}_{ac|bd}(P) = N_{ac}^T \, \mathrm{Flat}_{ac|bd}(Q) N_{bd}. \tag{3}$$

A third such expression, obtained in a similar way, is

$$\mathrm{Flat}_{ad|bc}(P) = N_{ad}^T \, \mathrm{Flat}_{ad|bc}(Q) N_{bc}. \tag{4}$$

The key observation underlying our proof of the identifiability of the tree topology for the (λ, κ)-state model (Theorem 2 and Corollary 3 below) is that, for generic parameter choices, the ranks of the matrices $\mathrm{Flat}_{ab|cd}(Q)$, $\mathrm{Flat}_{ac|bd}(Q)$, $\mathrm{Flat}_{ad|bc}(Q)$, are rather different, and this affects the ranks of the three corresponding flattenings of P. This will lead to explicit polynomials (i.e., *phylogenetic invariants*) whose vanishing can be used to identify the tree topology from which P arises, for generic parameters.

Theorem 2. *Suppose $\lambda < \kappa^2$. With P denoting a $\kappa \times \kappa \times \kappa \times \kappa$ tensor of indeterminants, let S_1, S_2, and S_3 denote the sets of $(\lambda+1)$-minors of the $\kappa^2 \times \kappa^2$ matrices $\mathrm{Flat}_{ab|cd}(P)$, $\mathrm{Flat}_{ac|bd}(P)$, and $\mathrm{Flat}_{ad|bc}(P)$, respectively. Let the varieties $Y_i = V(S_i) \subset \mathbb{C}^{\kappa^4}$ be their zero sets. Then*

(i) *$V_{T_i} \subseteq Y_j$ if, and only if, $i = j$.*
(ii) *If $P \in V_{T_1} \cup V_{T_2} \cup V_{T_3}$ and $P \in Y_i \smallsetminus (Y_j \cup Y_k)$ for distinct i, j, k, then $P \in V_{T_i} \smallsetminus (V_{T_j} \cup V_{T_k})$.*
(iii) *For distinct i, j, k, let $X_i = \phi_{T_i}^{-1}(Y_j \cup Y_k) \subsetneq \mathbb{C}^L$. Then X_i is a proper algebraic subvariety of the complex parameter space for the 4-taxon tree T_i, and for any parameters $s \in \mathbb{C}^L \smallsetminus X_i$ the tree T_i is identifiable from the joint distribution tensor $P = \phi_{T_i}(s)$ via the vanishing of the polynomials in S_i.*

Proof. Throughout, we may assume $i = 1, j = 2, k = 3$.

To establish (i), for any parameters $s \in \mathbb{C}^L$ on T_1, let $Q = Q(s)$, $P = P(s) = \phi_{T_1}(s)$, where Q and P are the tensors described above. Now $\mathrm{Flat}_{ab|cd}(Q)$ is a matrix of all zeros except for a single $\lambda \times \lambda$ submatrix whose entries are those of $A = \mathrm{diag}(\boldsymbol{\pi}_r) M_{rf}$. Thus

$$\mathrm{rank}(\mathrm{Flat}_{ab|cd}(Q)) = \mathrm{rank}(A) \leq \lambda.$$

By Equation (2), this implies $\mathrm{rank}(\mathrm{Flat}_{ab|cd}(P)) \leq \lambda$. Thus $\phi_{T_1}(\mathbb{C}^L) \subseteq Y_1$, and hence $V_{T_1} \subseteq Y_1$.

We now show $V_{T_1} \not\subseteq Y_2$ or Y_3, by finding an $s \in \mathbb{C}^L$ with $\phi_{T_1}(s) \notin Y_2 \cup Y_3$. To pick such an s, we choose each of $M_{ra}, M_{rb}, M_{fc}, M_{fd}$ to have the block form $(I_{\kappa\times\kappa}\ 0_{(\lambda-\kappa)\times\kappa})^T$. Then the Kronecker products $N_{ac}, N_{bd}, N_{ad}, N_{bc}$ all have block form

$$(I_{\kappa^2\times\kappa^2}\ 0_{(\lambda^2-\kappa^2)\times\kappa^2})^T.$$

Now choosing $\boldsymbol{\pi}_r$ and M_{rf} to have all positive entries, for instance, ensures that all entries of A are non-zero. Since $\mathrm{Flat}_{ac|bd}(Q)$ is a matrix of all zeros, except the entries of A which appear in the $((i,j),(i,j))$ positions, $\mathrm{Flat}_{ac|bd}(Q)$ is thus a diagonal matrix of rank λ^2. Similarly $\mathrm{Flat}_{ad|bc}(Q)$ is diagonal with full rank λ^2. Thus

$$\mathrm{Flat}_{ac|bd}(P) = N_{ac}^T \mathrm{Flat}_{ac|bd}(Q) N_{bd},$$
$$\mathrm{Flat}_{ad|bc}(P) = N_{ad}^T \mathrm{Flat}_{ad|bc}(Q) N_{bc},$$

both have rank $\kappa^2 > \lambda$ due to the particular form of $N_{ac}, N_{bd}, N_{ad}, N_{bc}$. Thus $\phi_{T_1}(s) \notin Y_2$ or Y_3.

Statement (ii) follows immediately from (i).

For (iii), note that the existence of the point s, constructed above, with $\phi_{T_1}(s) \notin Y_2$ or Y_3 shows X_1 is a proper subset of $\mathbb{C}^L$. That it is an algebraic variety follows from its definition as the zero set of all polynomials of the form $f \circ \phi_{T_1}$ where f vanishes on $Y_2 \cup Y_3$. □

Remark. The set X_i should be thought of as the set of 'bad' parameters for the model on T_i, for which this approach is unable to identify the tree topology from the resulting joint distribution. It is important that X_i be a proper subvariety of the parameter space since this immediately implies its dimension is smaller than that of the parameter space. If one restricts attention from complex to real parameters, or even to stochastic parameters, the points in X_i still form a set of lower dimension than the full parameter space. Thus for 'most' stochastic parameters, the topology is identifiable from the joint distribution.

Corollary 3. *The n-taxon bifurcating tree topology is identifiable for generic parameters of the (λ,κ)-state general Markov model when $\lambda < \kappa^2$. That is for each n-leaf tree T, there exists a proper subvariety X_T of the complex parameter space $\mathbb{C}^L$ such that the tree topology is identifiable from the joint distribution arising from any parameter choice $s \in \mathbb{C}^L \setminus X_T$ via the vanishing of certain explicit polynomials (to be described below).*

Proof. As is well known [SS03], to identify the topology of an n-leaf tree T, it is enough to identify the topology of each induced quartet tree relating four leaves of T.

Let $\mathbb{C}^L$ be the complex parameter space for the tree T, and let $\mathcal{Q}$ denote the collection of all 4-leaf trees induced by T. For each $T' \in \mathcal{Q}$, the parameter space for T' is $\mathbb{C}^{L'}$ and we have the following commutative diagram of polynomial maps:

$$\begin{array}{ccc} \mathbb{C}^L & \xrightarrow{\phi_T} & \mathbb{C}^{\kappa^n} \\ {\scriptstyle \alpha_{T'}}\downarrow & & \downarrow{\scriptstyle \mu_{T'}} \\ \mathbb{C}^{L'} & \xrightarrow{\phi_{T'}} & \mathbb{C}^{\kappa^4} \end{array}.$$

The map $\alpha_{T'}$ can be explicitly given by multiplication of matrix parameters for T to obtain matrix parameters for T', once a consistent choice of roots for T' and T is made. The map $\mu_{T'}$ is a marginalization map on tensors, where we sum over all but the 4 indices corresponding to leaves of T'.

For any $T' \in \mathcal{Q}$, identify its leaves with labels a, b, c, d so that T' is identified with T_1 of Theorem 2. Letting X_i, Y_i be the varieties defined in that theorem, consider varieties

$$X_T = \bigcup_{T' \in \mathcal{Q}} \alpha_{T'}^{-1}(X_1),$$
$$Y_T = \bigcap_{T' \in \mathcal{Q}} \mu_{T'}^{-1}(Y_1).$$

For any parameters $s \in \mathbb{C}^L \smallsetminus X_T$, $\phi_T(s) \in Y_T \smallsetminus \phi_T(X_T)$, and all 4-leaf induced tree topologies are identifiable by the vanishing of the polynomials defining Y_T. (An explicit set of polynomials defining Y_T can be taken to be the composition of the polynomials in S_1 of Theorem 2 with the marginalization maps $\mu_{T'}$.)

It only remains to show that X_T is a proper subvariety of $\mathbb{C}^L$. But it is easy to see that each of the maps $\alpha_{T'}$ is surjective. Therefore since $X_1 \subsetneq \mathbb{C}^{L'}$, for each $T' \in \mathcal{Q}$ we find $\alpha_{T'}^{-1}(X_1) \subsetneq \mathbb{C}^L$. Since X_T is a finite union of proper subvarieties of $\mathbb{C}^L$, we obtain $X_T \subsetneq \mathbb{C}^L$. □

We note that the idea of using rank conditions on flattenings of a data tensor to identify tree topology also appears in recent independent work of Eriksson [Eri05], where the SVD is used to give a novel algorithm for tree construction. That paper deals only with the general Markov model ($\lambda = \kappa$), and takes a slightly different approach to identifying splits for a tree without focusing on quartets.

Specializing our result to $\lambda = \kappa$, we recover the following result proved previously by Steel [Ste94] using the log-det distance, and then reproved by Eriksson.

Corollary 4. *The tree topology is identifiable for generic parameters in the κ-state general Markov model.*

5. Identifiability of tree topology for analytic (λ, κ)-state models

To deduce identifiability of the tree topology for analytic (λ, κ)-state submodels using Theorem 2 or Corollary 3 requires a little additional work, since, *a priori*, it is possible that the restricted parameters are not sufficiently generic to preserve identifiability.

Fix an analytic (λ, κ)-state model, and let $\psi : U \to \mathbb{C}^L$ denote its Markov map, giving parameters for the general (λ, κ)-state model in terms of parameters for the analytic model.

We first consider the identifiability of a 4-leaf tree topology, so let $T_1 = T_{ab|cd}$. For parameters u of the analytic model, we need only show that points $\psi(u)$ are generically not in the variety X_1 of Theorem 2.

With $I(X_1)$ denoting the ideal of all polynomials vanishing on X_1, consider the set

$$\widetilde{X}_1 = \psi^{-1}(X_1) = \{u : f \circ \psi(u) = 0, \text{ for all } f \in I(X_1)\} \subseteq U \subseteq \mathbb{R}^M.$$

Since ψ is an analytic map, so is $f \circ \psi$ for each polynomial f, and so $\widetilde{X}_1$ is an analytic subvariety of U. Thus if we establish that $\widetilde{X}_1$ is a proper subvariety of U, then generic points in U are mapped to generic points (off of X_1) in $\mathbb{C}^L$ by ψ. This establishes the following:

Lemma 5. *For $\lambda < \kappa^2$, consider an analytic (λ, κ)-state model on a 4-leaf tree, with parameter space U and Markov map ψ. If there is a single choice of parameters $u \in U$ such that $\psi(u) \notin X_1$, then the 4-leaf tree topology is identifiable for generic parameters.*

For ease of application to the specific models listed in Section 3, we deduce a weaker form of this.

Lemma 6. *Consider an analytic κ-base, m-class model on a 4-leaf tree. Suppose $m < \kappa$ and there is at least one choice of allowable parameters for which*

(i) *the Markov matrices for pendant edges are of the form $M_e = M_0 = (I_{\kappa\times\kappa}\ I_{\kappa\times\kappa} \dots\ I_{\kappa\times\kappa})^T$, and*
(ii) *if $\boldsymbol{\pi}_r$ is the root distribution and M_e the $\lambda \times \lambda$ Markov matrix assigned to the internal edge of the tree, then the $\kappa \times \kappa$ matrix*

$$N = M_0^T \operatorname{diag}(\boldsymbol{\pi}_r) M_e M_0$$

has at least $m\kappa + 1$ non-zero entries.

Then the tree topology is identifiable for generic parameters of the model.

Before proving this, we note that condition (i) means that no base changes occur on pendant edges, though class information is hidden. In condition (ii), N represents a joint distribution of bases, without class information, at the two internal nodes of the tree.

Proof. Since $m < \kappa$, then $m\kappa < \kappa^2$ and Lemma 5 applies.

For a parameter choice on the tree T_1 as described in the statement of the lemma, the joint distribution of bases at the leaves is given by P where

$$P(i,j,k,l) = \begin{cases} N(i,k) & \text{if } i=j,\ k=l \\ 0 & \text{otherwise} \end{cases}.$$

Therefore the matrices $\text{Flat}_{ac|bd}(P)$ and $\text{Flat}_{ad|bc}(P)$ are diagonal with at least $m\kappa+1$ non-zero entries. Hence they have rank at least $m\kappa+1$. This shows the parameters do not lie in X_1, and so the topology is identifiable for generic parameters. □

We now obtain the result that provided our original motivation for this work.

Corollary 7. *For the covarion model of Tuffley-Steel, if $\kappa \geq 3$ the 4-leaf tree topology is identifiable for generic parameters.*

Proof. This model is an analytic κ-base, 2-class model, and so we need $\kappa \geq 3$ to apply Lemma 6.

For any $R = (R_{ij}), \boldsymbol{\pi}, s_1, s_2$ with all $s_i, \pi_i, R_{ij} > 0$ for $i \neq j$, the matrix $\text{diag}(\boldsymbol{\pi}_r)\exp(Qt_e)$ has all positive entries as long as $t_e > 0$, so then the matrix N has all positive entries. Picking such parameters, with $t_e > 0$ for the internal edge of the tree, and $t_e = 0$ for all pendant edges, Lemma 6 gives the result. □

This result includes the $\kappa = 4$, 20 cases which apply to DNA and protein sequences. Note, however, that the identifiability of the tree topology for the $\kappa = 2$ covarion model remains an open question.

Finally, the result extends to trees with more than 4 leaves, by an argument analogous to that of Corollary 3.

Corollary 8. *For the covarion model of Tuffley-Steel, if $\kappa \geq 3$ then bifurcating tree topologies are identifiable for generic parameters.*

Though we omit the details, one similarly sees that tree topologies for the 4-base, m-class covarion model SSRV of [Gal01] are generically identifiable provided $m < 4$. Note that the implementation of the SSRV

in inference software described in that paper actually had $m = 4$, a case not covered by our theorem. It would of course be desirable to prove identifiability for that case, and larger m, as well.

Finally, we can apply this approach to non-covarion rate-variation models with a finite number of rate classes. As an example, we give the following result.

Corollary 9. *For the GM+GM+$\cdots$+GM model, with κ states and m classes where $m < \kappa$, bifurcating tree topologies are identifiable for generic parameters. In particular, when $\kappa = 4$, the tree topology is generically identifiable for the GM+GM+GM model.*

Proof. For the 4-leaf tree, consider any parameter choice where no substitutions occur on pendant edges in any of the classes, the root distribution has all positive entries, all Markov matrix entries are nonnegative, and for at least one class the $\kappa \times \kappa$ Markov matrix for that class on the internal edge has at least $\kappa+1$ positive entries. Then apply Lemma 6.

An argument analogous to that for Corollary 3 extends the result to trees with more leaves. □

Similarly, for the GTR+rate-classes model we obtain generic identifiability of tree topology provided the number of classes m is less than the number of bases κ. Note that while previous result on identifiability for this model [WS97, Rog01] have allowed a known continuous distribution of rates, they have also assumed a common rate matrix for all classes. Our result holds even for a model in which different classes have unrelated GTR rate matrices.

Finally, we note this approach proves generic identifiability of tree topologies for the GM+I model when $\kappa \geq 3$. However, for this particular model we will take a different approach in another paper [AR05a], obtaining identifiability for $\kappa \geq 2$ as well as some interesting explicit formulas for recovering proportions of invariable sites, and identifying other numerical parameters as well.

References

[AR03] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

[AR04] Elizabeth S. Allman and John A. Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *App. Math. Res. Express (AMRX)*, 2004:4:107–131, 2004.

[AR05a] Elizabeth S. Allman and John A. Rhodes. Identifying topologies and parameters for the GM+I model. 2005. in preparation.

[AR05b] Elizabeth S. Allman and John A. Rhodes. Phylogenetic ideals and varieties for the general Markov model, 2005. in preparation, `arXiv:math.AG/0410604`.

[AR05c] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 2005. to appear, `arXiv:q-bio.PE/0407035`.

[Baa98] Ellen Baake. What can and what cannot be inferred from pairwise sequence comparisons? *Math. Biosci.*, 154(1):1–21, 1998.

[BGP05] David Bryant, Nicolas Galtier, and Marie-Anne Poursat. Likelihood caclulation in molecular phylogenetics. In Olivier Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 33–62. Oxford University Press, 2005.

[Bun71] Peter Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*, pages 387–395, Edinburgh, 1971. Edinburgh University Press.

[CF87] James A. Cavender and Joseph Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

[CGS05] Marta Casanellas, Luis David Garcia, and Seth Sullivant. Catalog of small trees. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 291–304. Cambridge University Press, 2005.

[CH91] J. T. Chang and J. A. Hartigan. Reconstruction of evolutionary trees from pairwise disbributions on current species. In E. M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 254–257. Interface Foundation, 1991.

[Cha96] Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[CHHP00] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Bio. and Evol.*, 17:1529–1541, 2000.

[CKS03] B. Chor, A. Khetan, and S. Snir. Maximum likelihood on four taxa phylogenetic trees: Analytic solutions. *RECOMB'03*, 2003.

[CS05] Marta Casanellas and Seth Sullivant. The strand symmetric model. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 305–321. Cambridge University Press, 2005.

[Eri05] Nicholas Eriksson. Tree construction using singular value decomposition. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*, pages 347–358. Cambridge University Press, 2005.

[ERSS04] Nicholas Eriksson, Kristian Ranestad, Bernd Sturmfels, and Seth Sullivant. Phylogenetic algebraic geometry. 2004. to appear, in proceedings of "Projective Varieties with Unexpected Properties," Siena, Italy, `arXiv:math.AG/0407033`.

[ES93] Steven N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[EW04] Steven N. Evans and Tandy Warnow. Unidentifiable divergence times in rates-across-sites models. 2004. `arXiv:q-bio.PE/0408011`.

[FM70] Walter M. Fitch and Etan Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4:579–593, 1970.

[Gal01] Nicolas Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18(5):866–873, 2001.

[Gas05] Olivier Gascuel, editor. *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford, 2005.

[Hen89] Michael D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38:310–321, 1989.

[Lak87] J.A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

[PMCH01] David Penny, Bennet J. McComish, Michael A. Charleston, and Michael D. Hendy. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *J. Mol. Evol.*, 53:711–723, 2001.

[PS05] Lior Pachter and Bernd Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge, 2005.

[Rog01] James S. Rogers. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.*, 50(5):713–722, 2001.

[SHP98] Mike Steel, Michael D. Hendy, and David Penny. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. *Discrete Appl. Math.*, 88(1-3):367–396, 1998.

[SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[SS05] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12(2):204–228, 2005. `arXiv:q-bio.PE/0402015`.

[SSE93] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

[SSH94] M.A. Steel, L. Székely, and M.D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

[Ste94] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Letters*, 7(2):19–23, 1994.

[TS98] Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147(1):63–91, 1998.

[WS97] Peter J. Waddell and M.A. Steel. General time-reversible distances with unequal rates across sites: Mixing Γ and inverse Gaussian distributions with invariant sites. *Mol. Phylo. Evol.*, 8(3):398–414, 1997.

ESA: Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775

E-mail address: `e.allman@uaf.edu`

JAR: Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775
E-mail address: j.rhodes@uaf.edu